An integrative approach to drug repositioning: a use case for semantic web technologies

Student: Paul Rigor
Mentor: Olivier Bodenreider

July 29, 2011

Abstract

This report describes my summer rotation project at the Lister Hill National Center for Biomedical Communications in the National Library of Medicine at the National Institutes of Health. My project introduced me to key technologies and trends in computer aided drug repositioning[24] and the emerging role of the semantic web[5]. This proof of concept illustrates the potential utility of semantic web technologies and datasets from the Linked Open Drug Data (LODD)[15] task force which is coordinated by the Health Care and Life Sciences (HCLS) Interest Group from the World Wide Web (W3C) Consortium. The academic entities from this group have provided a readily integrate-able and interoperable set of drug data which I have reused in an attempt to replicate the discoveries of Campillos, et. al in 2008. These resources highlight an improvement in scientific communication; however, they are not without their caveats[4, 15, 23].

Introduction

In the past decade, there have been several paradigm shifts with respect to medicine, drug discovery, and the technologies that seek to integrate the deluge of heterogeneous data in the health care and life sciences. Specifically, the growth of biomedical, pharmaceutical, and genomic data has provided opportunities for accelerating translational medicine, permitting the discovery of novel targets for existing drugs, thus providing a low attrition, cost-effective drug development path compared to *de novo* methods [3].

For example, as a poster child of this drug repositioning trend, Thalidomide has spent almost six decades in the spot light where it has been marked by both its tragic beginnings in the 1950's and its serendipitously rediscovered beneficial applications from the mid 1960's up to the present day. Indeed, this drug which was discontinued because of its teratogenic effects, now has arisen to become a potential cancer therapeutic[29].

As shown in Figure 1 and reviewed by Ashburn and Thor[3], drug repositioning has offered a faster time-to-market by avoiding the initial phase of drug discovery which includes target discovery, lead

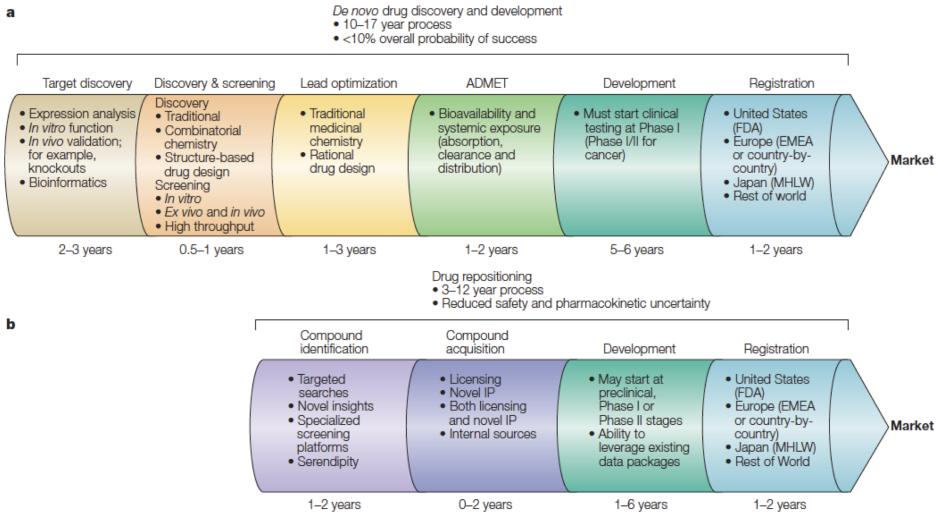

Figure 1: An overview of *de novo* and drug repositioning pipelines [Ashburn and Thor 2004].

discovery, and lead optimization. Further mitigating the cost of resources, existing drugs have already passed ADMET (absorption, distribution, metabolism, excretion, and toxicity) tests and clinical trials thereby reducing "safety and pharmacokinetic uncertainty." Some caveats still exist, however. For example, new clinical tests must be developed and approved for the proposed novel indication. In the case for viagra, standard metrics had to be developed for normative erection in order to distinguish from dysfunctional erection.

Indeed, drug repositioning has paved the way for rational drug design. Instead of phenotypic assays that involve trial-and-error, we are delving deeper into the biological processes (and components) that underlie disease. Moreover, one aspect of drug repositioning hinges on the understanding of the so-called polypharmacology of drugs[17]; that is, as suggested by the myriad of drug side effects, there exist unintended and hidden targets for existing drugs[6].

Along with *in vivo* and *in vitro* (including serendipitous) methods, several *in silico* tools, databases and frameworks have been employed by pharmaceutical companies and academic researchers alike to discover novel applications of existing drugs. These methods include the structure- and ligand-based virtual screening[28, 12], target binding-site similarity[20], network pharmacology[2, 16, 13], and pharmacogenomics[21, 14].

For our present study, our proof of concept hinged on the similarity of side effects among drugs as described by Campillos, et. al. For example, in some instances, the molecular targets of marketed drugs are known and annotated. However, not all FDA-approved drugs have known and experimentally validated targets.

Thus, our goal was to infer targets of drugs which have not been annotated. In order to achieve this, we focused on existing FDA-approved drugs which are annotated in DrugBank[18] as well as the database of side effects which are annotated in SIDER[19]. Further, we postulate that the 'unknown' target of one drug can be inferred from the 'known' target of another drug by the

similarity between their side effects. Moreover, we can also infer additional targets for drugs with already known targets.

Because our strategy relied on the integration of the drug information from DrugBank and side effect information SIDER, we utilized several semantic web technologies and off-the-shelf resources to mitigate the initial barrier to the integration process when dealing with traditional web services and relational databases.

Materials and Methods

Data sources

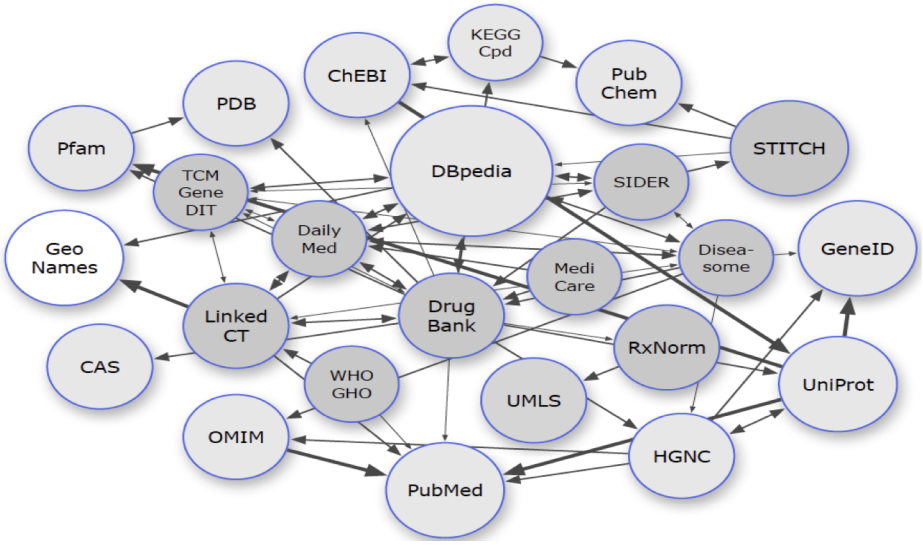

Figure 2: A map of the data sets curated by the LODD task force.

We have deployed several instances of the open-source version of Virtuoso [27] and have populated them with the LODD[15] datasets including Uniprot[9], Diseasome[11], SIDER[19], ClinicalTrials[8], DailyMed[10], and DrugBank[18]. The Virtuoso software suite provided both the triple store for the aforementioned linked and RDF[26] data along with the SPARQL[25] endpoints to these resources. Aside from the aforementioned RDFized datasets, we also loaded and evaluated over 40 other RDFized datasets from the life science domain: Bio2RDF[1] and Chem2Bio2RDF[7].

For our proof of concept, we have selected the DrugBank and SIDER datasets. DrugBank contains information on the known drug targets for marketed drugs, while SIDER contains side effect information for those marketed drugs. A caveat, which will be discussed in the Discussion section, is the currency of the datasets. At the time of use, the LODD curated SIDER data contained only 275 drugs with annotated side effects, whereas the most recent version of SIDER contains 888 marketed drugs with annotated side effects.

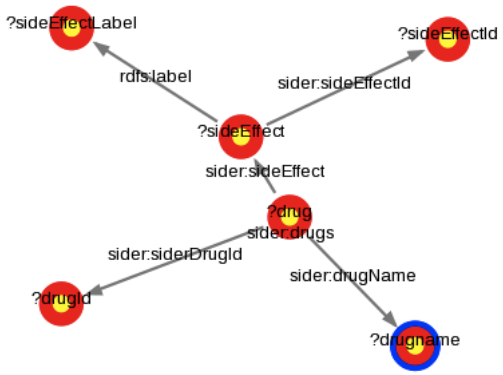

Figure 3: A graph layout of the SPARQL query.

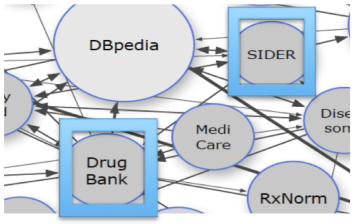

Figure 4: The subset of the LODD datasets that we eventually used: SIDER and DrugBank

Integration

To integrate the SIDER and DrugBank datasets, we utilized the SPARQL graph query language to interrogate the Virtuoso instances, which served as the storage and access points (SPARQL endpoint) for the SIDER and DrugBank LODD datasets. SPARQL is a superset of the Structured Query Language (SQL) used in relational database management systems and is also the standard query language for the semantic web.

We interrogated the endpoints with the SPARQL query below, which generated a merged graph of drug-to-side effect and drug-to-target interactions. The core of the query rests inside the *WHERE* clause which provides several graph patterns to match against the data graphs of the LODD datasets. More importantly, the key graph pattern involves the *owl:sameAs* predicate (line 16) which provided the critical semantic link that matched the drug entities from the SIDER and DrugBank data graphs.

```
1  PREFIX owl: <http://www.w3.org/2002/07/owl#>
2  PREFIX sider: <http://www4.wiwiss.fu-berlin.de/sider/resource/sider/>
3  PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
4  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5  SELECT *
6  FROM <http://semanticweb.ics.uci.edu/LODD/SIDER#>
7  FROM <http://semanticweb.ics.uci.edu/LODD/DrugBank#>
8  WHERE {
9   ?drug a sider:drugs ;
10          sider:drugName ?drugname ;
11          sider:siderDrugId ?drugId ;
12          sider:sideEffect ?sideEffect .
13  ?sideEffect rdfs:label ?sideEffectLabel ;
14          sider:sideEffectId ?sideEffectId .
15  ?ddDrug rdf:type drugbank:drugs;
16    owl:sameAs ?drug .
17    OPTIONAL { ?ddDrug drugbank:target ?target }
18  }
```

We subsequently transformed the resulting graph into two feature sets for subsequent similarity analyses. One feature set contained a vector of side effects per drug. Another contained a vector of targets per drug. Thus, for our analysis two similarity matrices were generated for each feature set.

Analysis

We calculated the pair-wise drug-drug side effect similarity using the Jaccard similarity metric,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

which computes the ratio of the intersection of side effects to the union of side effects. We also calculated the same similarity matrix for drug target similarity. We performed a chi-square test to determine the correlation between side effect similarity and shared target(s). We created several contingency tables by varying the Jaccard side effect similarity score (from 0.1 to 1.0) in order to determine the most significant side effect similarity score.

Results

Figure 5 illustrates that the majority of the SIDER-annotated drugs share very few side effects.

As shown on figures 6 and 7, we were able to capture the intuitive relationship between drugs and their shared targets based on their shared side effects. With the chi-square test, we validated the significance of this relationship at a Jaccard side effect similarity score of 0.5 (p-value = 5.797201e-209, degrees of freedom = 1, chi-square statistic = 951). Further, Table 1 shows the grouping we captured for similar classes of drugs: antihistamines and barbiturates. Thus, we have established that side effect similarity is a significant predictor of drugs sharing the same targets.

Drug 1	Drug 2	Side effect similarity score
Carbinoxamine	Diphenhydramine	0.90
Chlorthalidone	hydroflumethiazide	0.90
Carbinoxamine	Dexchlorpheniramine Maleate	0.87
Carbinoxamine	Clemastine	0.86
Dexchlorpheniramine Maleate	Diphenhydramine	0.82
Clemastine	Diphenhydramine	0.80
Pentobarbital	secobarbital	0.86

Table 1: A list of drugs that have high similarity score and are annotated to share the same targets.

We have further validated our hypothesis, as shown on Table 2, by capturing the grouping of drugs with similar side effects but unannotated targets. These drugs have actually been annotated as being in the same class of drugs according to DrugBank drug categories.

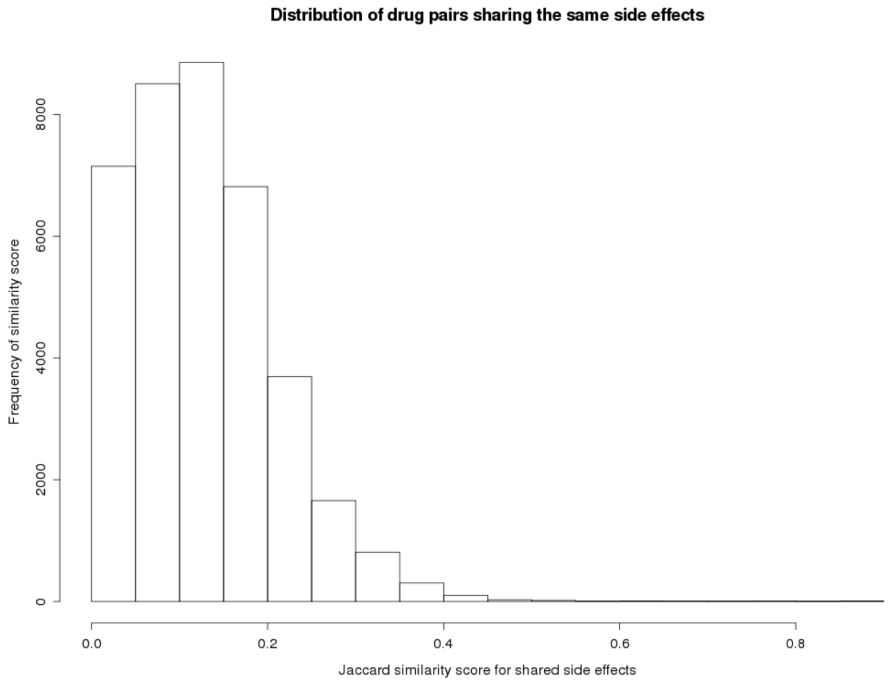

Figure 5: Distribution of Jaccard similarity score for side effects.

Discussion and Conclusions

Although our proof-of-concept validated side effect similarity as an indicator for drugs sharing the same target, because of the limitations of our datasets, we were unable to capture a more intriguing relationship: discovering shared targets between drugs from different drug categories. As tables 1 and 2 illustrate, we have only captured the grouping of drugs within a similar class. In fact (data not shown), even using the drug category similarity (as defined by DrugBank) rendered similar results. Perhaps we needed a more discriminant set of drug classification (e.g., NDFRT[22])as well as an updated RDF version of the SIDER database in order uncover yet known drug targets.

Challenges

As the abstract alluded to, there are still caveats to using semantic web technologies and resources. With respect to the ease of integration, utilizing off-the-shelf datasets from the semantic web does present an attractive alternative to traditional integration of heterogenous datasets. However, semantic web technologies and standards are still evolving – from the underlying triple store and the graph representation, to the domain-specific ontologies and the SPARQL graph query language. Also, the lack of a standardized ontology among Bio2RDF, Chem2Bio2RDF, and LODD still presents a hurdle to the wide adoption in the health sciences.

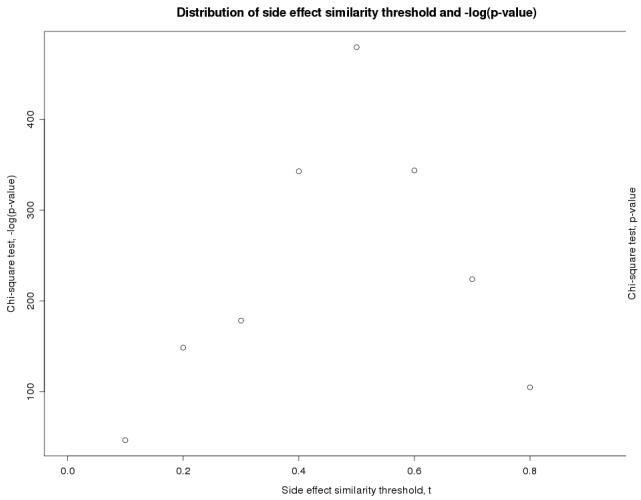

Figure 6: Distribution of -log(p-value) from a chi-square test of independence for pair-wise drug-drug side effect similarity and shared targets.

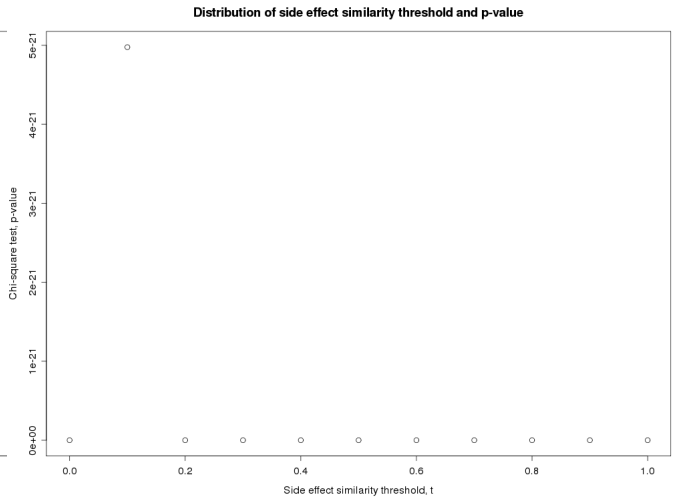

Figure 7: Distribution of p-values from a chi-square test of independence for pair-wise drug-drug side effect similarity and shared targets.

Drug 1	Drug 2	Side effect similarity score
Bendroflumethiazide	Chlorthalidone	0.67
Risperidone	Citalopram Hydrobromide	0.54
propantheline	Orphenadrine Citrate	0.54
Fluvoxamine Maleate	Gabapentin	0.52
Citalopram Hydrobromide	Topiramate	0.52
Risperidone	Topiramate	0.51
Citalopram Hydrobromide	Ropinirole Hydrochloride	0.51

Table 2: A list of drugs that have significant side effect similarity score but are annotated not to share the same targets.

Furthermore, gaps still exists between the motivations of original data providers and the goals of data curators. As such, as consumers of semantic web data, we are still at the mercy of this evolving technology. Moreover, as we have experienced, the currency of the data still precludes the promise of a systematized approach in the automated mining and inference of interesting relationships among the molecular players and processes underlying diseases.

My experience and future work

Although still evolving, semantic web technologies present an exciting frontier in which to integrate existing projects and resources. In particular, participation with an international community such as the BioRDF and LODD task forces have not only broadened my approach in computer-aided drug discovery and repositioning, but have also introduced several useful resources at the National

Library of Medicine.

Acknowledgements

NLM Mentor: Dr. Olivier Bodenreider; Ph.D. Advisor: Dr. Pierre Baldi; Jonathan Mortensen; May Cheh, Summer Rotation Program; My colleagues; Institute for Genomics and Bioinformatics, Donald Bren School for Information and Computer Science, University of California in Irvine; Lister Hill National Center for Biomedical Communications; National Library of Medecines Biomedical Informatics Training Program; ORISE; National Institutes of Health

References

[1] Peter Ansell. Model and prototype for querying multiple linked scientific datasets. *Future Generation Computer Systems*, 27(3):329 – 333, 2011.

[2] DK Arrell and A Terzic. Network systems bilogy for drug discovery. *Nature Clinical Pharmacology and Therapeutics*, 88, 2010.

[3] Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 2004.

[4] Christine Chichester Peter-Bram't Hoen Johan T den Dunnen Gertjan van Ommen Erik van Mulligen Bharat Singh Rob Hooft Marco Roos Joel Hammond Bruce Kiesel Belinda Giardine Jan Velterop Paul Groth Erik Schultes Barend Mons, Herman van Haagen. The value of data. *Nature Genetics*, 2011.

[5] Tim Berners-Lee. http://www.w3.org/designissues/semantic.html.

[6] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[7] Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David Wild. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, 11(1):255+, May 2010.

[8] clinicaltrials.gov. http://clinicaltrials.gov/.

[9] The UniProt Consortium. Ongoing and future developments at the universal protein resource. *Nucleic Acids Res*, 39, 2011.

[10] dailymed.nlm.nih.gov. http://dailymed.nlm.nih.gov/dailymed.

[11] Diseasome.org. http://www.diseasome.org/.

[12] Michael J. Keiser et. al. Predicting new molecular targets for known drugs. *Nature*, 462, 2009.

[13] Muhammed A Yldrm et. al. Drug-target network. *Nature Biotechnology*, 25:1119 – 1126, 2007.

[14] Yamanishi et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26:246–254, 2010.

[15] Linked Open Drug Data Task Force. http://www.w3.org/wiki/hclsig/lodd.

[16] Andrew L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 2008.

[17] Andrew L. Hopkins. Drug discovery: Predicting promiscuity. *Nature*, 2009.

[18] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. Drugbank 3.0: a comprehensive resource for omics research on drugs. *Nucleic Acids Research*, 39(suppl 1):D1035–D1041, 2011.

[19] Letunic I Jensen LJ-Bork P Kuhn M, Campillos M. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 2010.

[20] Lei Xie Philip E. Bourne Li Xie, Jerry Li. Drug discovery using chemical systems biology: Identification of the protein-ligand binding network to explain the side effets of cetp inhibitors. *PLOS: Computational Biology*, 5(5), 2009.

[21] Daniel L. Rubin Katrina L. Easton Joshua M. Stuart Russ B. Altman Micheal Hewett, Diane E. Oliver and Teri E. Klein. Pharmgkb: the pharmacogenetics knowledge base. *Nucleic Acids Res*, 30, 2002.

[22] NDFRT UMLS NLM NIH. http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ndfrt/. 2011.

[23] Bodenreider O. Ontologies and data integration in biomedicine: Success stories and challenging issues. *In: Bairoch A, Cohen-Boulakia S, Froidevaux C, editors. Proceedings of the Fifth International Workshop on Data Integration in the Life Sciences (DILS 2008*, 2008.

[24] Tudor I. Oprea, Sonny Kim Nielsen, Oleg Ursu, Jeremy J. Yang, Olivier Taboureau, Stephen L. Mathias, Irene Kouskoumvekaki, Larry A. Sklar, and Cristian G. Bologa. Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Molecular Informatics*, 30(2-3):100–111, 2011.

[25] SPARL query language for RDF. http://www.w3.org/tr/rdf-sparql-query/. 2011.

[26] Resource Description Framework (RDF). http://www.w3.org/rdf/. 2011.

[27] OpenLink Software. Virtuoso suite: http://www.openlinksw.com/.

[28] Steven L. Swann, Scott P. Brown, Steven W. Muchmore, Hetal Patel, Philip Merta, John Locklear, and Philip J. Hajduk. A unified, probabilistic framework for structure- and ligand-based virtual screening. *Journal of Medicinal Chemistry*, 54(5):1223–1232, 2011.

[29] Steve K. Teo, David I. Stirling, and Jerome B. Zeldis. Thalidomide as a novel therapeutic agent: new uses for an old product. *Drug Discovery Today*, 10(2):107 – 114, 2005.